תיאור סט הנתונים UCI Heart Disease Data :

מקור הנתונים: הdataset-נלקח מ :Kaggle

קישור: https://www.kaggle.com/datasets/redwankarimsony/heart-disease-data

תיאור סט הנתונים:

סט הנתונים מכיל נתונים על מטופלים שמטרתם לחזות האם יש למטופל מחלת לב. המידע נוגע למאפיינים קליניים שונים שנאספו על כל מטופל, וכולל משתנים מגוונים שיכולים להיות רלוונטיים בניבוי נוכחות מחלת לב.

ניסוח הבעיה:

מחלות לב הן אחת מהסיבות המובילות לתמותה בעולם, ניבוי מוקדם של מחלה זו יכול להציל חיים באמצעות טיפול מונע. על בסיס סט הנתונים של חולי לב, המטרה היא לזהות ולחזות את הסיכון למחלות לב על סמך מאפיינים קליניים כמו גיל, מין, סוג כאב בחזה, לחץ דם, רמות כולסטרול ועוד. נשאלת השאלה: *האם ניתן לנבא נוכחות מחלת לב אצל מטופלים בהתבסס על מאפיינים קליניים קיימים?*

שאלת המחקר:

אילו משתנים קליניים (כגון גיל, לחץ דם, רמות כולסטרול, תעוקת חזה במאמץ) הם החזקים ביותר בחיזוי סיכון למחלת לב?

השערות:

משתנים קליניים כמו גיל, רמות כולסטרול, תעוקת חזה במאמץ וכולסטרול יהיו בעלי השפעה מובהקת על הסיכון למחלת לב. כלומר, משתנים אלו יהיו בעלי כוח חיזוי גבוה יותר בהשוואה למשתנים אחרים.

ציפיות:

המודלים יזהו את המשתנים הקליניים המרכזיים הקשורים באופן חזק לסיכון למחלת לב. אני מצפה לראות תוצאה גבוהה של משמעות סטטיסטית עבור משתנים כמו גיל, לחץ דם, רמות כולסטרול ותעוקת חזה במאמץ.

EDA:

השדות שסט-הנתונים מכיל:

הdataset מכיל 921 רשומות (שורות) ו-16 שדות (עמודות עיקריות.)

השדות וסוג הנתונים(data types) :

1. Id: מספר סידורי שלם של מטופלים
2. Age: גיל המטופל (מספר שלם, נומרי).
3. Sex: מין המטופל (קטגוריה, 0 = אישה, 1 = גבר).
4. cp (chest pain type): סוג כאב החזה (קטגוריה, 0-3).
5. trestbps (resting blood pressure): לחץ דם במנוחה (מספר שלם, נומרי).
6. chol (serum cholesterol): רמת כולסטרול בדם (מספר שלם, נומרי).
7. Dataset: ארץ מגורים של הנבדק .

8. fbs (fasting blood sugar): סוכר בצום (קטגוריה, 1 אם > 120 mg/dl, אחרת 0).
9. restecg (resting ECG results): תוצאות ECG במנוחה (קטגוריה, (קטגוריה, 0= normal, 1= lv hypertrophy, 3= st-t abnormality).
10. thalach (maximum heart rate achieved): דופק מרבי שהושג (מספר שלם, נומרי).
11. exang (exercise induced angina): אנגינה שנגרמה כתוצאה מפעילות גופנית (קטגוריה, 1 =כן, 0 =לא).
12. Oldpeak: דיכאון מקטע ST יחסי למנוחה (מספר עשרוני, נומרי).
13. slope (slope of the peak exercise ST segment): שיפוע מקטע ST במאמץ (קטגוריה, 0-2).
14. ca (number of major vessels colored by fluoroscopy): מספר כלי דם עיקריים שנצבעו (קטגוריה, 0-4.
15. thal (thalassemia): תלסמיה (קטגוריה, 1-3).
16. Target: תוצאה (קטגוריה, 0 = לא חולה לב, 1 = חולה לב, 2 = חולה לב בינוני, 3 = חולה לב חמור, 4 = חולה לב חמור מאוד).

פעולות הניקיון / השלמה / נרמול שבוצעו:

<u>ניקיון נתונים:</u>

- מחיקת עמודות לא רלוונטיות: מחקתי את עמודות ה-`id` וה-`dataset`, שלא תורמות למודל.
- מיפוי נתונים קטגוריאליים לערכים מספריים: עמודות קטגוריאליות (כמו מין, כאב חזה, בדיקת סוכר,...) עברו המרה לערכים מספריים כדי שניתן יהיה להשתמש בהם במודלים.

<u>השלמת נתונים חסרים:</u>

- מילוי נתונים חסרים עם החציון: השתמשתי ב-`SimpleImputer` כדי למלא ערכים חסרים בעמודות מספריות באמצעות החציון (שיטה מומלצת להימנע מהשפעה של ערכים קיצוניים).

<u>נרמול נתונים:</u>

- ביצעתי נרמול סטנדרטי (Standardization) על עמודות מספריות באמצעות `StandardScaler`. הנרמול חשוב כדי להבטיח שכל המאפיינים יהיו באותה סקאלה.

<u>האם נדרשנו לאחד מספר מקורות נתונים?</u>

במקרה זה, לא נדרש חיבור של מספר datasets .כל הנתונים הנחוצים לניתוח מגיעים ממקור יחיד.

חישוב מטריקות, המטריקות שחישבתי:

1. Precision (דיוק):

- מדד המגדיר את אחוז התחזיות הנכונות מתוך כל התחזיות החיוביות.
- חישוב: $Precision = \frac{TP}{TP+FP}$ *TP: תחזיות חיוביות נכונות, *FP: תחזיות חיוביות שגויות.
- מתוך כל המקרים שהמודל חזה כמחלת לב (קטגוריה 1), כמה היו באמת אנשים עם מחלת לב.

2. Recall (זיהוי):

- מדד המראה את היכולת של המודל לזהות נכון את הדוגמאות החיוביות מתוך כלל הדוגמאות החיוביות.
- מתוך כל האנשים שבאמת יש להם מחלת לב, כמה מהם המודל זיהה בצורה נכונה.

3. F1-Score:

- ממוצע הרמוני בין Precision ל-Recall. משמש למדוד את הדיוק הכולל של המודל.
- נותן מדד מאוזן כאשר יש לך הבדל גדול בין Precision לRecall

4.ROC AUC (Receiver Operating Characteristic - Area Under Curve) :

- המדד ליכולת המודל להבחין בין מחלקות חיוביות ושליליות.
- חישוב: העקומה ROC מציגה את היחס בין True Positive Rate ל-False Positive Rate בערכים שונים של סף החלטה. המדד AUC מחושב על ידי חישוב השטח תחת עקומת ROC.
- ערך קרוב ל-1 מצביע על כך שהמודל מצליח להבחין בצורה טובה בין הקטגוריות (0 ו-1), בעוד שערך קרוב ל-0.5 מצביע על ניחוש אקראי.

בכל אחת מהמטריקות הללו, נעשה שימוש על התחזיות שהמודל הפיק בהשוואה לנתוני האמת (y_test) כדי להעריך את ביצועיו.

התוצאות שהתקבל:

לאחר ביצוע הניקוי, ההשלמה והנרמול של הנתונים, ביצעתי אימון והערכה על מספר מודלים. התוצאות מתייחסות לדיוק המודלים השונים לפי מדדים כמו Precision, Recall, F1-Score, ו-ROC AUC. הנה התוצאות והתובנות מכל מודל:

ROC AUC	Accuracy	F1 Score	Recall	Precision	קטגוריה	מודל
0.819	0.76	0.75	0.73	0.77	0	Logistic Regression
		0.76	0.78	0.75	1	
0.919	0.82	0.82	0.79	0.84	0	Random Forest Classifier
		0.83	0.86	0.81	1	
0.869	0.81	0.80	0.78	0.82	0	Gradient Boosting Classifier
		0.81	0.83	0.79	1	
0.904	0.8	0.79	0.74	0.84	0	SVM (Support Vector Machine)
		0.81	0.86	0.77	1	

תובנות:

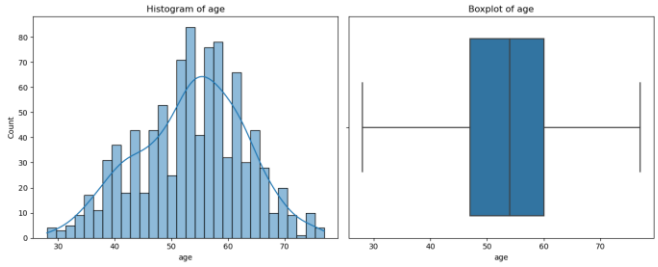
- Logistic Regression: המודל הביא לתוצאות מאוזנות עם דיוק וזיהוי סבירים לשתי הקטגוריות, וערך ROC AUC מעל 0.8 מצביע על הפרדה טובה בין המחלקות.
- Random Forest Classifier: מודל זה השיג את הביצועים הטובים ביותר עם ערכי דיוק וזיהוי גבוהים במיוחד ו-ROC AUC קרוב ל-0.92, מה שמעיד על כך שהמודל מבצע הפרדה מצוינת בין הקטגוריות.
- Gradient Boosting Classifier: תוצאות המודל היו טובות מאוד, אם כי מעט נמוכות מ-Random Forest, עם ROC AUC קרוב ל-0.87.
- SVC: המודל הציג ביצועים חזקים עם דיוק כללי של 0.80 וערך ROC AUC קרוב ל-0.90, מה שמעיד על יכולת הפרדה טובה.

לפי התוצאות, המודל Random Forest Classifier הוא העדיף מבין המודלים שנבדקו. הנה סיבות לכך:

1. דיוק גבוה: המודל השיג את ערכי ה-Precision וה-Recall הגבוהים ביותר עבור שתי הקטגוריות (0.84 עבור קטגוריה 0 ו-0.81 עבור קטגוריה 1).
2. F1 Score: ה-F1 Score עבור שתי הקטגוריות היה גבוה (0.82 עבור קטגוריה 0 ו-0.83 עבור קטגוריה 1), מה שמעיד על איזון טוב בין דיוק וזיהוי.
3. ROC AUC: ערך ה-ROC AUC היה הגבוה ביותר (0.919), מה שמעיד על כך שהמודל מבצע הפרדה מצוינת בין הקטגוריות.
4. דיוק כללי(Accuracy): מודל זה הציג דיוק כללי גבוה של 0.82, שהוא מהגבוהים בין כל המודלים שנבדקו.

- Univariate Analysis

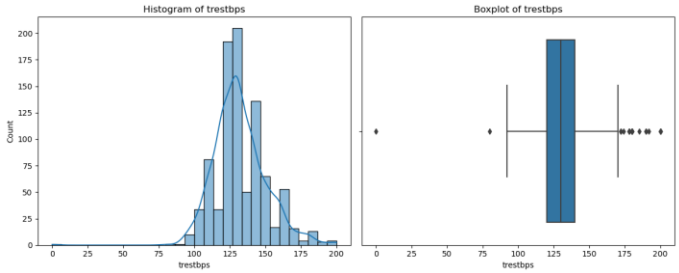
תובנות על ההתפלגות של **משתנים נומריים** מתחת לויזואליזציות:



Summary Statistics of age

Statistic	Value
count	920.0
mean	53.51086956521739
std	9.424685209576857
min	28.0
25%	47.0
50%	54.0
75%	60.0
max	77.0

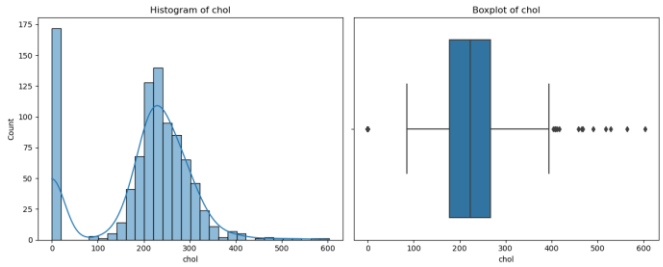
Age - גיל: ההתפלגות נראית נורמלית בערך, עם ריכוז גבוה של גילאים סביב 50-60, כפי שנראה גם בהיסטוגרמה וגם ב boxplot. אין ערכים חריגים (outliers) בולטים, שכן הגבולות של תיבת החציון (boxplot) מקיפים את רוב הערכים, ואין חריגים מחוץ לטווח.



Summary Statistics of trestbps

Statistic	Value
count	920.0
mean	131.99565217391304
std	18.451299974216376
min	0.0
25%	120.0
50%	130.0
75%	140.0
max	200.0

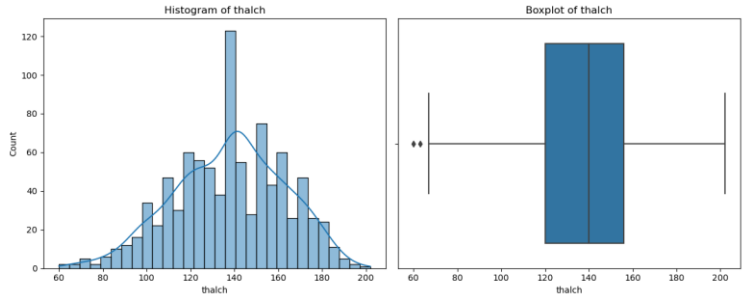
לחץ הדם במנוחה - (trestbps): ההתפלגות של לחץ הדם במנוחה (trestbps) נוטה לנורמלית עם ערכים מרוכזים בין 120 ל-140, והשיא סביב 130. קיימים מספר ערכים חריגים (outliers) מתחת ל-75 ומעל 175, כפי שנראה ב-boxplot.



Summary Statistics of chol

Statistic	Value
count	920.0
mean	199.90869565217392
std	109.04017054571457
min	0.0
25%	177.75
50%	223.0
75%	267.0
max	603.0

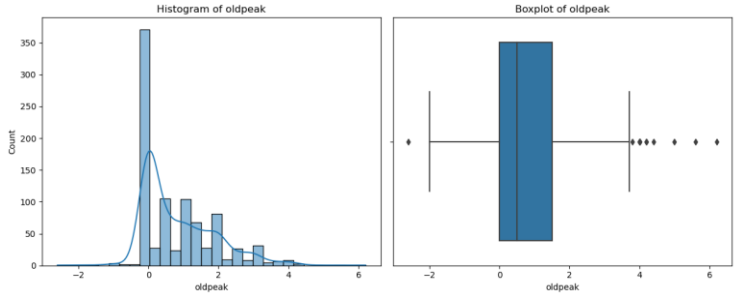
chol - רמת כולסטרול בדם: ההתפלגות של ערך הכולסטרול (chol) מוטה ימינה, עם ריכוז גבוה של ערכים סביב 200-250 כפי שנראה בהיסטוגרמה. קיימים ערכים חריגים (outliers) בצד הימני של התפלגות הכולסטרול, כפי שנראה בתרשים תיבת החציון, עם כמה ערכים החורגים מעל 400.



Summary Statistics of thalch

Statistic	Value
count	920.0
mean	137.69239130434784
std	25.145235299317232
min	60.0
25%	120.0
50%	140.0
75%	156.0
max	202.0

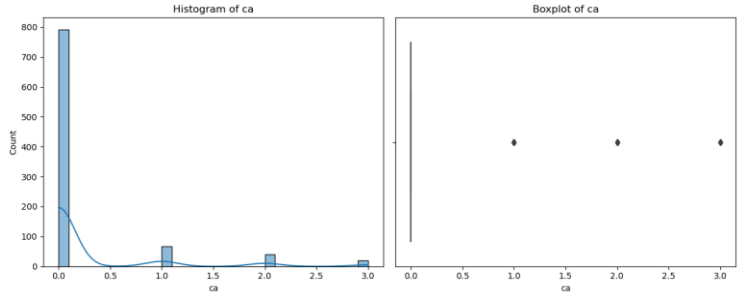
thalach - דופק מרבי שהושג: ההתפלגות של הערך "thalach" היא נורמלית פחות או יותר, עם ריכוז גבוה של ערכים בטווח 120-160, כפי שנראה בהיסטוגרמה. קיימים ערכים חריגים נמוכים מתחת ל-80, כפי שנראה בתרשים תיבת החציון.



Summary Statistics of oldpeak

Statistic	Value
count	920.0
mean	0.8532608695652173
std	1.0580486984557564
min	-2.6
25%	0.0
50%	0.5
75%	1.5
max	6.2

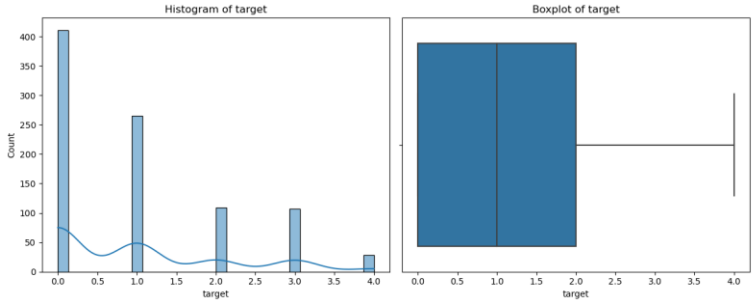
Oldpeak - דיכאון מקטע ST יחסי למנוחה: הערך של oldpeak מתפלג בצורה חיובית עם רוב הערכים קרובים ל-0. קיימים מספר ערכים גבוהים יותר שנראים חריגים, במיוחד מעבר לערך 4 (נמצאים מחוץ לתחום הקופסה).



Summary Statistics of ca

Statistic	Value
count	920.0
mean	0.2271739130434782 5
std	0.6289364843815157
min	0.0
25%	0.0
50%	0.0
75%	0.0
max	3.0

ca - מספר כלי דם עיקריים שנצבעו: ההתפלגות של ca היא קטגוריאלית עם ערכים שלמים, כאשר רוב הדגימות נמצאות בערך 0. ישנם כמה ערכים חריגים, במיוחד הערך 3, המהווה נקודת חריג בקופסה.

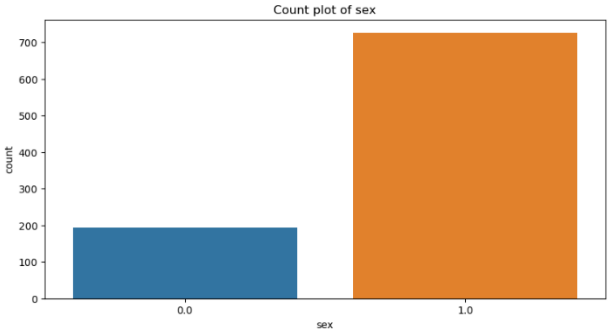


Summary Statistics of target

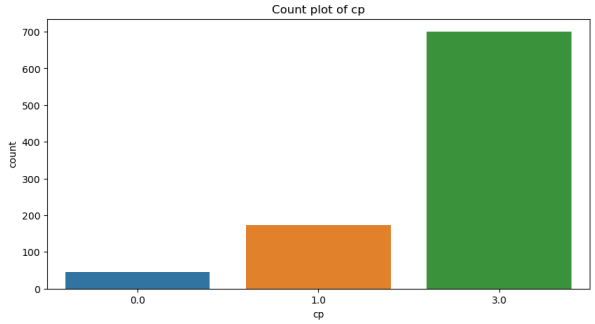
Statistic	Value
count	920.0
mean	0.9956521739130435
std	1.1426934033380 34
min	0.0
25%	0.0
50%	1.0
75%	2.0
max	4.0

Target: גם כאן ההתפלגות היא קטגוריאלית, עם ערכים מ-0 עד 1. רוב הערכים הם 0, וישנן מספר דגימות בערכים. אין משתנים קטגוריים חריגים בקופסה.

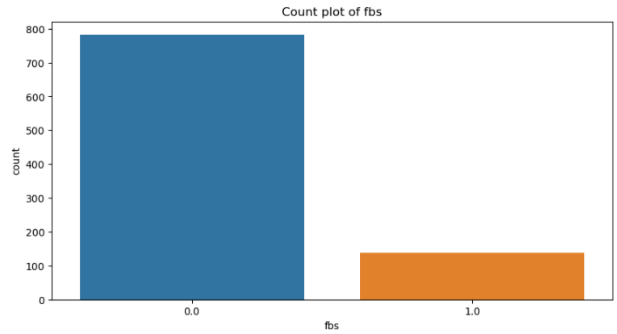
משתנים קטגוריאליים:



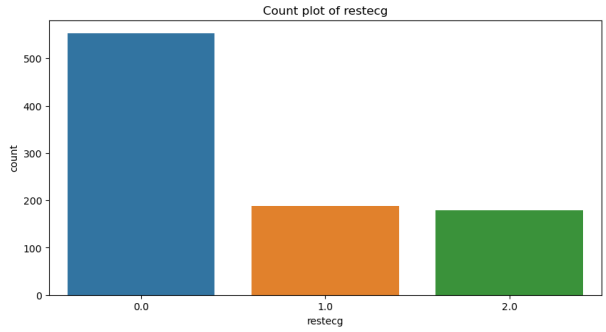
מגדר: בדגימה יש קרוב ל-200 נשים וכמעט 700 גברים. זו התפלגות לא מאוזנת מבחינת מגדר, כאשר מספר הגברים גבוה משמעותית ממספר הנשים.



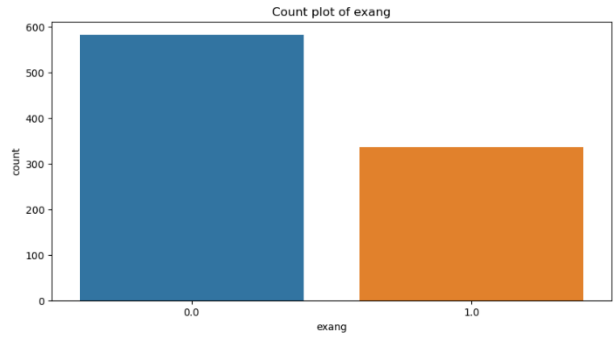
Cp-סוג כאב חזה: רוב המקרים הם מסוג 3.0 (כ-700 מקרים), עם מספר קטן יחסית של מקרים מסוג0 ו-1. נראה כי כאב חזה מסוג 3.0 נפוץ בהרבה מהשאר, אך לא נראים ערכים חריגים יוצאי דופן.



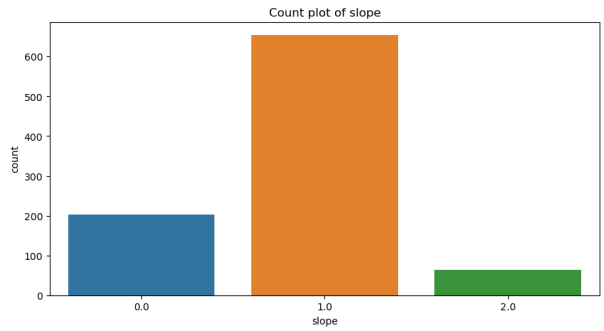
סוכר בצום Fbs: ניתן לראות שקרוב ל-100 מקרים עם ערך 1 (כלומר רמת סוכר גבוהה מ-120 mg/dl), בעוד שבכ-800 מקרים רמת הסוכר נמוכה מ-120 mg/dl (ערך 0). רוב הנבדקים בדגימה זו לא סובלים מסוכר גבוה בצום.



תוצאות ECG במנוחה -Restecg:הגרף מראה שיותר מ-500 מקרים נמצאו תקינים (0), כ-200 מקרים הראו היפרטרופיה של החדר השמאלי (1), וקצת פחות מ-200 מקרים הראו חריגות ST-T (2). כלומר, רוב המקרים מצביעים על ECG תקין.



אנגינה שנגרמה כתוצאה מפעילות גופנית- Exang: יותר מ-300 מקרים היו עם אנגינה בעקבות פעילות גופנית (ערך 1), בעוד שקצת פחות מ-600 מקרים לא סבלו מאנגינה לאחר פעילות גופנית (ערך 0). כלומר, רוב הנבדקים לא חוו אנגינה כתוצאה מפעילות גופנית.

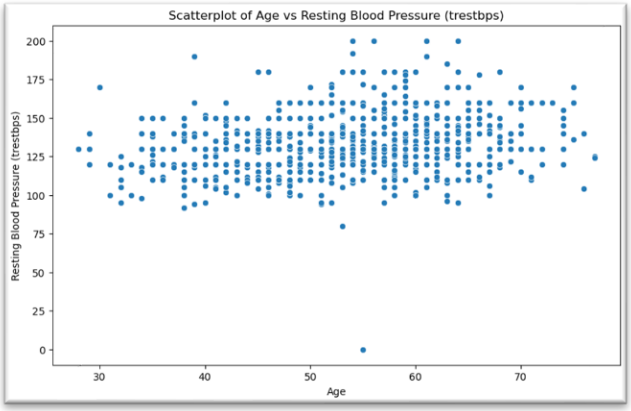


שיפוע מקטע ST במאמץ -Slope: רוב המקרים מצביעים על שיפוע 1 במקטע ST בזמן מאמץ.

סיכום: מהגרפים ניתן לראות שיש שונות בין המשתנים השונים, אך בחלקם יש דומיננטיות לקטגוריות מסוימות, כמו יותר גברים מאשר נשים, רמות סוכר תקינות ברוב האוכלוסייה, ושיפוע ST נפוץ בקטגוריה 1.

Bivariate Analysis, ויזואליזציות עם תובנות:

א. <u>תרשים הפיזור מציג את הקשר בין גיל (Age) ללחץ דם במנוחה (Resting Blood Pressure - trestbps):</u>



<u>התפלגות כללית</u>:

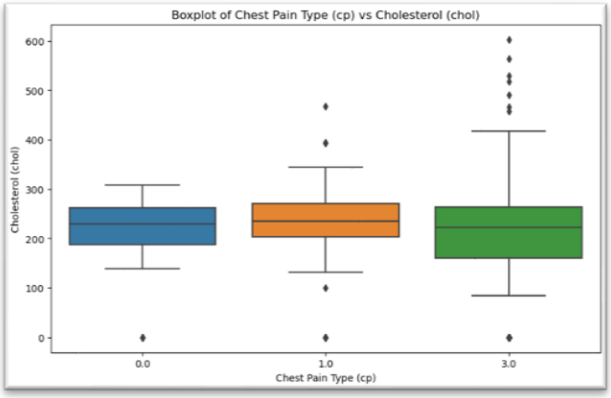
- הערכים נעים בעיקר בין גילאים 30 ל-70.
- רוב ערכי לחץ הדם נעים בטווח שבין 100 ל-175 יחידות.
- רוב הנקודות מתרכזות בין לחץ דם 125 ל-150 יחידות, ללא קשר ברור לגיל.

<u>ערכים חריגים</u>:

- יש מספר ערכים חריגים בלחץ הדם, בעיקר ערכים נמוכים מאוד (כמו ערך שמתקרב לאפס) המופיעים במספר נקודות בודדות.
- ערכים אלו עשויים להיות תוצאה של מדידה שגויה או נתונים שגויים ויש לבדוק אותם.

<u>סיכום</u>: לא נראה שיש קשר מובהק בין גיל ללחץ הדם במנוחה, אבל יש ערכים שמצביעים על נתונים חריגים.

ב. ***תרשים Boxplot מציג את ההשוואה בין סוגי כאב בחזה (cp) לבין רמות הכולסטרול (chol):***



התפלגות כללית:

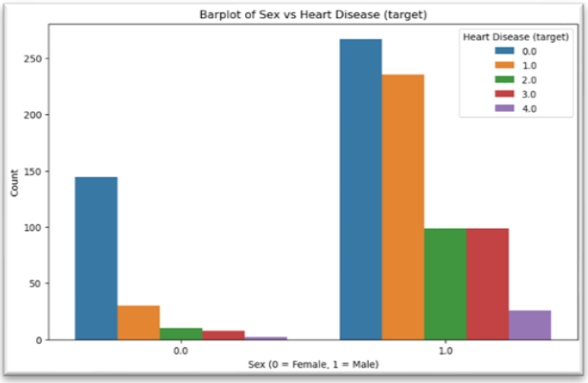
- לכל סוגי כאב החזה (0, 1, 3) טווחים דומים ברמות הכולסטרול, עם ריכוז גבוה סביב 200-300.
- הערכים המרכזיים (חציון, IQR) בכל הקבוצות נמצאים קרוב זה לזה, כאשר החציון נע סביב 250 יחידות כולסטרול בכל הקבוצות.

ערכים חריגים:

- יש מספר ערכים חריגים בכל הקבוצות, כאשר בקבוצות cp=0 ו-cp=1 יש ערכים חריגים נמוכים מאוד (כמעט אפס).
- בקבוצה cp=3 יש ריכוז של ערכים חריגים גבוהים יותר, מעל 400 ואף מעל 500 יחידות כולסטרול, מה שמצביע על פיזור גבוה יותר בקטגוריה זו.

סיכום: לא נראה שיש הבדל מובהק בין הקבוצות ברמות הכולסטרול, כי יש מספר ערכים חריגים בכל קבוצה.

ג. ***הגרף מציג את ההתפלגות של מין (נשים=0, גברים=1) מול מחלת לב (target):***



נשים(Sex = 0):

- מרבית הנשים (מעל 150) שייכות לקטגוריה 0 (ללא מחלת לב).
- יש מעט מאוד נשים שיש להן דרגות מחלת לב גבוהות יותר, כאשר הדרגה הגבוהה ביותר (4) כמעט ולא מיוצגת.

גברים (Sex = 1):

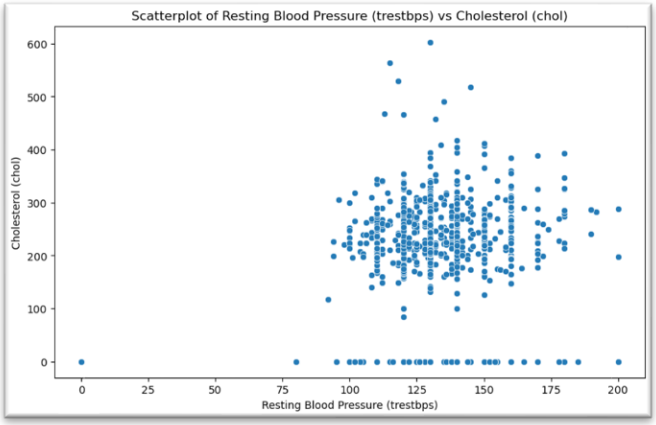
- רוב הגברים שייכים לקטגוריה 0 (ללא מחלת לב) וגם לקטגוריה 1 (דרגת מחלה קלה), כאשר יש מספר רב של גברים בכל אחת משתי הקטגוריות האלו.
- יש ייצוג ניכר של גברים עם דרגות מחלה 2 ו-3, אך דרגת המחלה 4 פחות נפוצה.

התפלגות חריגה:

- אצל נשים יש הבדל משמעותי בין הימצאות מחלת לב בהשוואה לגברים. הרבה יותר נשים נמצאות ללא מחלת לב (0) לעומת גברים, בעוד שיותר גברים נמצאים בדרגות חומרה גבוהות יותר.

סיכום: הגרף מדגים באופן ברור שגברים נוטים יותר לסבול ממחלת לב לעומת נשים, ובעיקר בדרגות חומרה גבוהות יותר, אך רוב הנשים והגברים ללא מחלת לב נמצאים בדרגת חומרה 0.

תרשים פיזור (Scatterplot) של לחץ דם במנוחה (trestbps) מול כולסטרול (chol):



היקף הנתונים המרכזי:

- מרבית הערכים של ***trestbps*** נעים בין 100-150.
- רוב הערכים של כולסטרול נעים בין 200 ל-350.
- רוב הנקודות מתרכזות בטווח הזה, מה שמצביע על כך שזהו הטווח הנפוץ ביותר עבור הנתונים הנמדדים.

ערכים חריגים:

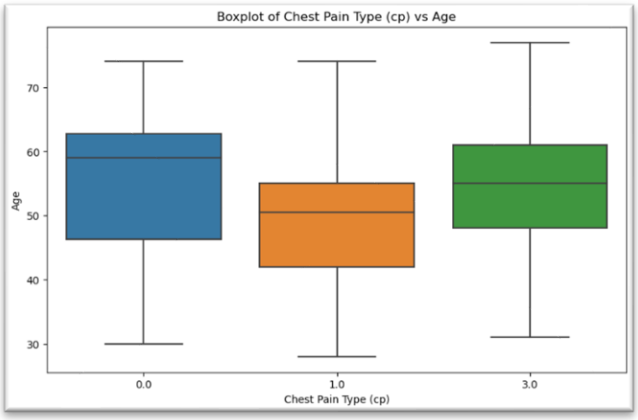
- יש מספר נקודות עם ערכי כולסטרול גבוהים מעל 500, שהן ערכים חריגים.
- יש ערכים חריגים מאוד נמוכים של כולסטרול שמתקרבים לאפס, מה שלא נפוץ מבחינה ביולוגית תקינה.
- קיימים ערכי לחץ דם חריגים, מעל 175, כמו גם ערכי לחץ דם מאוד נמוכים שמתקרבים לאפס.

התפלגות כללית:

- לא ניתן לזהות קשר לינארי ברור בין לחץ דם לכולסטרול. הנתונים מתפזרים בצורה רחבה ואקראית.
- ייתכן שיש גורמים נוספים שמשפיעים על שניהם, ולאמדובר בקורלציה ישירה.

סיכום: * אין קשר ברור וישיר בין לחץ דם במנוחה לרמת הכולסטרול על פי הגרף.
* קיימים ערכים חריגים בולטים גם בלחץ הדם וגם בכולסטרול, שיכולים להיות מעניינים לבדיקה נוספת.

ד. ***גרף המוצג הוא תרשים Boxplot המתאר את הקשר בין סוג הכאב בחזה (cp) לגיל (Age):***



סוג כאב בחזה 0:
- הגיל החציוני עבור כאב cp הוא סביב 60.
- טווח הגילאים נע בין כ-35 לכ-75.
- קופסת הנתונים מצביעה על כך שרוב האנשים עם סוג זה של כאב בחזה נמצאים בטווח גילאים של 50 עד 65.

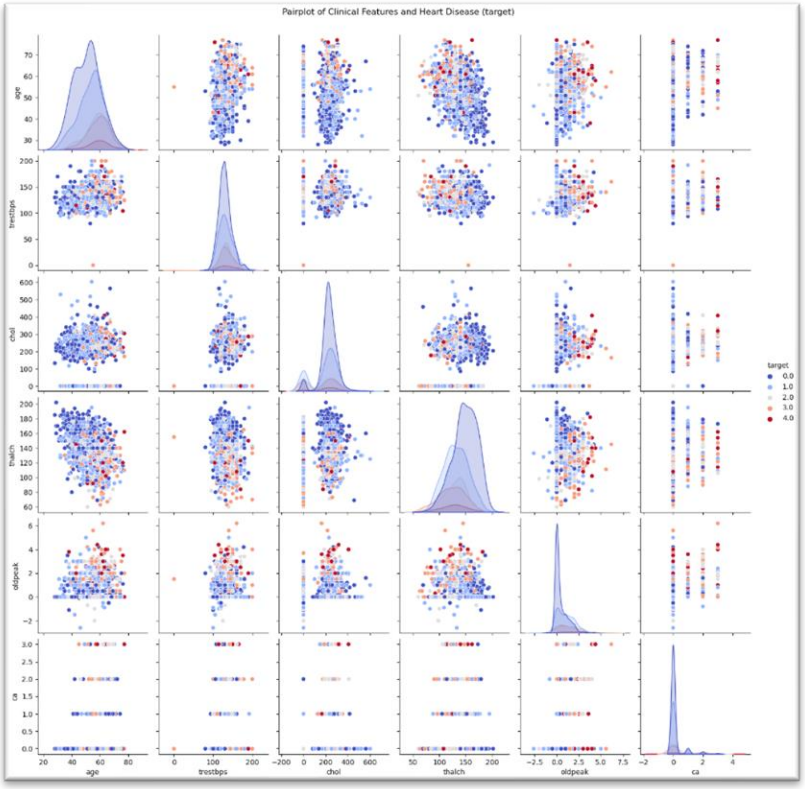
סוג כאב בחזה 1:
- הגיל החציוני עבור סוג זה של כאב נמוך יותר בהשוואה לשאר, ועומד על כ-50.
- טווח הגילאים רחב, נע בין גיל 30 - 75.
- הקופסה צרה יחסית, ומרבית האנשים נמצאים בטווח של גילאי 45 עד 60.

סוג כאב בחזה 3:
- הגיל החציוני עבור סוג זה של כאב הוא סביב 55.
- טווח הגילאים נע בין 30 לכ-75.
- הנתונים מפוזרים באופן רחב יחסית בטווח של 45 עד 65.

סיכום: * יש הבדל בגיל החציוני בין סוגי הכאב בחזה, כאשר סוג 1 מתאפיין בגילאים נמוכים יותר באופן יחסי.
* טווח הגילאים רחב בכל אחד מסוגי הכאב, אך ניתן לראות שסוגי כאב 0 ו-3 נוטים להיות נפוצים יותר אצל אנשים מבוגרים יותר, בעוד סוג 1 נפוץ יותר בגילאים צעירים.

Multivariate Analysis

ו. גרף Pairplot המציג את ההתפלגויות והקשרים בין משתנים קליניים שונים: גיל, לחץ דם במנוחה, רמות כולסטרול, קצב לב מרבי, ירידת ST במאמץ, מספר כלי דם עיקריים שצבעם נצבע, ואת מחלת הלב (target).



תובנות מהגרף:

1. התפלגות של כל משתנה בנפרד:
- גיל: מתפלג בגאוסיאן (פעמון) כשרוב המשתתפים בגילאים 40-60.
- לחץ דם במנוחה (trestbps): מתפלג עם ערכים שבין 100 ל-200, עם ריכוז גבוה יותר סביב 120-140.
- רמות כולסטרול (chol): ערכים נעים בין 100 ל-600, עם ריכוז גבוה סביב 200-300.
- קצב לב מרבי (thalach): מתפלג בגאוסיאן, עם רוב הערכים בין 100 ל-200 ושיא בסביבות 150-160.
- ירידת ST במאמץ (oldpeak): רוב הערכים נמוכים מ-2, עם מעט ערכים קיצוניים.
- מספר כלי דם עיקריים (ca): רוב הנתונים נמצאים בין 0 ל-3, אך יש גם כמה ערכים קיצוניים עד 4.

2. קשרים בין המשתנים:
- גיל (age) וכולסטרול (chol): אין קשר ברור מאוד.
- קצב לב מרבי (thalach) וירידת ST במאמץ (oldpeak): יש נטייה לערכים גבוהים יותר של ירידת ST במאמץ כאשר קצב הלב המרבי נמוך יותר, מה שעשוי לרמוז על קשר אפשרי.

- ca ו-oldpeak (ירידת ST במאמץ): ישנן קבוצות מוגדרות היטב לערכים מסוימים של שני המשתנים, מה שעשוי להצביע על קשר כלשהו בין המדדים הללו לבין סיכון למחלת לב.

3. קשרים עם target:

- thalach ו-target: קשר הפוך ברור, כאשר ערכים נמוכים של קצב לב מרבי מקושרים ל-target גבוה יותר (סיכון גבוה יותר למחלת לב).
- ירידתST במאמץ ו-target:ערכים גבוהים של ירידת ST במאמץ קשורים עם ערכים גבוהים יותר של target.
- ca (מספר כלי דם עיקריים שצבעם נצבע) ו-target: נראה שיש קשר חזק בין ca ל-target, כאשר ערכים גבוהים יותר של ca (כמו 3 או 4) נוטים להיות מקושרים עם ערכים גבוהים יותר של target.

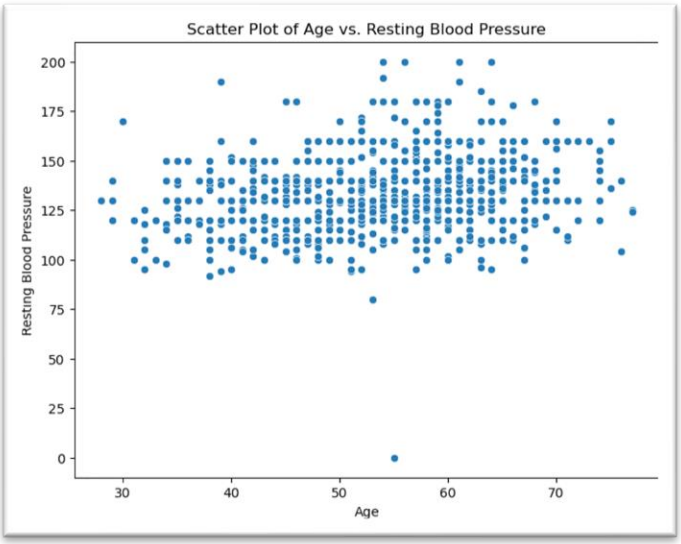
4. ערכים חריגים:

- chol (כולסטרול): יש כמה ערכים חריגים (מעל 500), אשר יוצאים מחוץ לטווח הנפוץ.
- oldpeak (ירידת ST במאמץ): ישנם כמה ערכים גבוהים במיוחד (מעל 5), שהם חריגים.
- ca: נראה שיש ריכוז מוגבל של ערכים, אך ערך 4 הוא די חריג ומייצג מקרים נדירים יותר.

<u>סיכום</u>: הגרף מצביע על כך שהמשתנים thalach (קצב לב מרבי), oldpeak (ירידת ST במאמץ), ו-ca (מספר כלי דם עיקריים שצבעם נצבע) הם המשתנים החזקים ביותר בקשר שלהם למחלת לב. קיימים גם ערכים חריגים במיוחד ב-chol (כולסטרול), oldpeak ו-ca, אשר עשויים להוות נתוני קצה (outliers).

ויזואליזציות מרכזיות נוספות:

<u>ז. הגרף מראה פיזור של גיל מול לחץ דם במנוחה:</u>



<u>פיזור כללי</u>: ניתן לראות שהנתונים מרוכזים בעיקר בין גילאי 40 ל-70, עם לחץ דם שנע בין 100 ל-175 מ"מ. הפיזור יחסית אחיד בין גילאים אלו, כאשר רוב המדידות נמצאות בטווח 120-160.

<u>ערכים חריגים:</u>

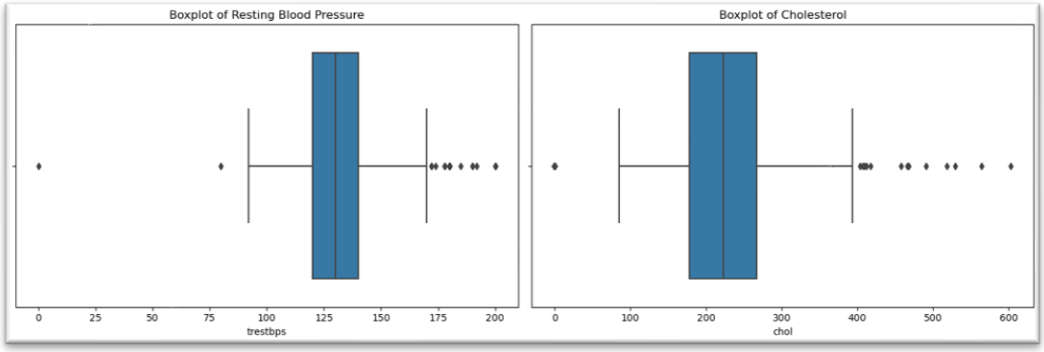
- ישנם ערכים חריגים נמוכים מאוד של לחץ דם, מתחת ל-75, כאשר יש אפילו נקודה בודדת עם לחץ דם של 0. זהו כנראה ערך חריג שנובע מטעות במדידה.
- יש מספר נקודות עם לחץ דם גבוה מ-200, שנחשבות כערכים חריגים.

<u>מגמת הקשר</u>: נראה שאין מגמה ברורה של עלייה או ירידה בלחץ הדם עם העלייה בגיל. הנתונים מראים פיזור רחב בכל הגילאים, ולכן לא ניתן להסיק קשר חזק בין גיל ללחץ דם במנוחה מתוך הגרף הזה.

<u>סיכום</u>: התפלגות הנתונים נראית די הומוגנית בין גילאי 40 ל-70, אך חשוב לשים לב לערכים החריגים שהתגלו בלחץ הדם.

<u>ח. הגרף מכיל שני תרשימי Boxplot, אחד עבור לחץ דם במנוחה (trestbps) והשני עבור רמת כולסטרול (chol).</u>

<u>הנה תובנות מהגרף:</u>



1. Boxplot של לחץ דם במנוחה (trestbps):

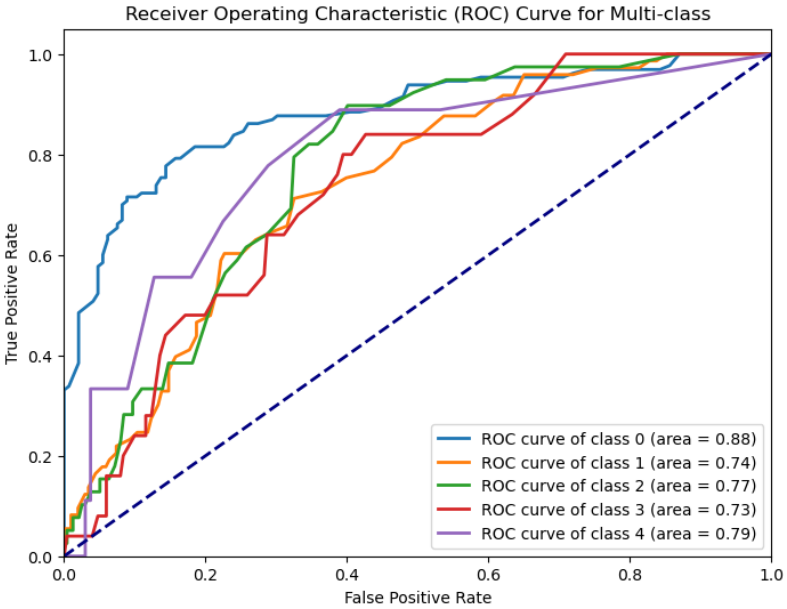
- התפלגות מרכזית: רוב הנתונים נעים בטווח של כ-120 - 140 מ"מ כספית, עם חציון שנראה באזור 130.
- ערכים חריגים: ישנם כמה ערכים חריגים נמוכים (בערך 0 ו-75 מ"מ כספית), וכן ערכים חריגים גבוהים מעל 175. הערך הקיצוני ביותר קרוב ל-200.

2. Boxplot של כולסטרול (chol):

- התפלגות מרכזית: רוב הנתונים נעים בטווח של 200 עד 300 מ"ג/ד"ל, עם חציון שנראה באזור 240.
- ערכים חריגים: ישנם מספר ערכים חריגים מעל 400, והערך החריג ביותר מתקרב ל-600. כמו כן, יש ערך חריג נמוך מאוד, שנראה כ-0, שעשוי להיות תוצאה שגויה או חסרה.

סיכום: שני התרשימים מראים התפלגות די תקינה עם מספר ערכים חריגים בקצוות, במיוחד בערכי לחץ דם נמוכים וערכי כולסטרול גבוהים. הערכים החריגים דורשים תשומת לב נוספת, שכן הם עשויים להיות תוצאה של טעויות במדידה או להצביע על מצבים חריגים בריאותיים.

ט. עקומות ROC לניבוי רב-מחלקתי של מחלת לב:



1. זיהוי טוב למחלקה 0 (לא חולה לב): המודל מזהה היטב אנשים ללא מחלת לב, עם AUC של 0.88.

2. דיוק בינוני בזיהוי מצבי מחלה שונים: המחלקות עם מחלת לב (1-4) זוהו בדיוק נמוך יותר, עם AUC שנע בין 0.73 ל-0.79.

3. זיהוי טוב יותר בקצוות: המודל מצליח יותר לזהות מצבים קיצוניים – ללא מחלה או מחלה חמורה מאוד – לעומת מצבי הביניים.

4. המודל מתאים לסינון ראשוני, עם זיהוי ראשוני של מקרים בריאים ומסוכנים, ומפנה להמשך בירור.

ריכוז התובנות וכתיבת הנרטיב:

קהל היעד:

הקהל שצפוי להקשיב לסיפור זה ולהביט בויזואליזציות מורכב מאנשי מקצוע רפואיים, חוקרים, ומקבלי החלטות במערכת הבריאות. בנוסף, ל חוקרי דאטה ומדעני נתונים המעוניינים להבין את המודלים לחיזוי מחלות לב ואת היכולות שלהם בניבוי סיכונים רפואיים.

מסר מרכזי:

הסיפור מתמקד בחשיבות של ניתוח נתונים רפואיים כדי לזהות מאפיינים קליניים מרכזיים המשפיעים על הסיכון למחלות לב. המטרה היא להדגיש כיצד ניתן להשתמש במודלים חישוביים כדי לנבא מחלת לב על בסיס מאפיינים קליניים, ולהציג את היתרונות של מודלים מסוימים בחיזוי נכון ומדויק יותר.

תובנות מרכזיות מהדאטה:

1. חשיבות משתנים קליניים: מאפיינים כמו גיל, לחץ דם, סוג כאב בחזה ורמות כולסטרול נמצאו כמשפיעים עיקריים על הסיכון למחלת לב.

2. חלוקה מגדרית: הנתונים מראים הטיה מובהקת במגדר, כאשר יותר גברים נבדקו, וזה עשוי להשפיע על תוצאות המודלים.
3. דיוק המודלים: המודל Random Forest השיג את הביצועים הטובים ביותר בחיזוי הסיכון למחלת לב, עם תוצאות ROC AUC גבוהות, מה שמצביע על יכולת הפרדה טובה בין חולים ללא חולים.

התובנות המרכזיות מהדאטה המבוססות על ניתוח נתוני מחלת הלב:

- <u>גיל וקשר למחלת לב</u>: רוב הנתונים מתרכזים בגילאי 40-60, עם קשר אפשרי לגורמי סיכון כגון לחץ דם ורמות כולסטרול. עם זאת, אין קשר ברור בין גיל ללחץ דם במנוחה.
- <u>לחץ דם במנוחה</u>: ההתפלגות נעה בין 100 ל-200 מ"מ כספית, עם ערכים חריגים בשני הקצוות. ערכים חריגים נמוכים עשויים לנבוע מטעויות במדידה, והערכים הגבוהים מעל 175 עלולים להוות גורם סיכון משמעותי.
- <u>רמות כולסטרול</u>: רוב המדדים נמצאים בטווח הנפוץ של 200-300 מ"ג/ד"ל, אך ישנם ערכים חריגים מעל 400, המצביעים על מצבים בריאותיים חריגים.
- <u>קצב לב מרבי</u>: קצב לב מרבי (thalach) נמוך נמצא בקשר הפוך למחלת לב, כאשר ערכים נמוכים יותר קשורים לעלייה בסיכון למחלת לב.
- <u>ירידת ST במאמץ</u>: ישנו קשר ברור בין ירידת ST במאמץ (oldpeak) לבין הסיכון למחלת לב. ערכים גבוהים יותר קשורים לעלייה בסיכון למחלה.
- <u>מספר כלי דם (ca)</u>: זהו משתנה מפתח בקשר למחלת לב. ככל שמספר כלי הדם העיקריים שנצבעו (ca) עולה, כך עולה גם הסיכון למחלת לב, במיוחד בערכים גבוהים של 3 ו-4.
- <u>מחלקות והמודל (ROC)</u>: מחלקה 0 (ללא מחלת לב) זוהתה ברמה גבוהה מאוד, בעוד שמחלקות 1-4, המייצגות דרגות שונות של מחלת לב, הראו ביצועים בינוניים, מה שמצביע על צורך בשיפור הזיהוי במחלקות אלה.
- <u>ערכים חריגים (Outliers)</u>: במספר משתנים, כמו כולסטרול (chol), לחץ דם במנוחה (trestbps), וירידת ST במאמץ (oldpeak), נמצאו ערכים חריגים, העלולים להשפיע על דיוק המודל ודורשים התייחסות נוספת.

סיכום: המשתנים המרכזיים המשפיעים על הסיכון למחלת לב כוללים את מספר כלי הדם העיקריים שנצבעו (ca), קצב לב מרבי (thalach), וירידת ST במאמץ (oldpeak). ישנם ערכים חריגים הדורשים בדיקה נוספת, והזדמנויות לשיפור המודל בזיהוי המחלקות השונות.

התשובה לשאלת המחקר:

בהתבסס על הוויזואליזציות שבוצעו, התשובה לשאלת המחקר "אילו משתנים קליניים הם החזקים ביותר בחיזוי סיכון למחלת לב?" היא:

1. <u>מספר כלי דם עיקריים שנצבעו (ca)</u> – נמצא כמשתנה החזק ביותר בחיזוי הסיכון למחלת לב. ככל שמספר כלי הדם שנצבעו גבוה יותר (במיוחד ערכים כמו 3 ו-4), כך עולה הסיכון למחלת לב.
2. <u>קצב לב מרבי (thalach)</u> – קשר הפוך ברור נמצא בין קצב לב מרבי לסיכון למחלת לב. קצב לב מרבי נמוך קשור בסיכון גבוה יותר למחלה.
3. <u>ירידת ST במאמץ (oldpeak)</u> – ערכים גבוהים של ירידת ST במאמץ מקושרים באופן מובהק לסיכון מוגבר למחלת לב.

משתנים אלו הראו קשרים מובהקים למחלת הלב, על פי הוויזואליזציות שנערכו. משתנים נוספים, כמו גיל, לחץ דם ורמות כולסטרול, הציגו קשרים פחות חזקים בהשוואה לשלושת המשתנים העיקריים הללו.

זה נקבע בעיקר על סמך גרף ה-Pairplot (גרף ו') והקשרים שנמצאו בו בין המשתנים הקליניים למחלת הלב (target). בנוסף, גם גרפים אחרים תרמו להבנת התובנות:

Pairplot: הציג את הקשרים בין משתנים קליניים כמו קצב לב מרבי (thalach), ירידת ST במאמץ (oldpeak), מספר כלי דם עיקריים שנצבעו (ca), וגיל לבין הסיכון למחלת לב (target). בגרף זה התגלו הקשרים החזקים ביותר עבור משתנים כמו ca, thalach, ו-oldpeak:

גרף ה-ROC: הדגים את ביצועי המודל בניבוי מחלת לב והראה שמחלקה 0 זוהתה טוב יותר (AUC גבוה), אך הקשרים שנמצאו ב-Pairplot נתמכו גם בגרף זה בכך שהם המשתנים שהכי תרמו לניבוי.

הוויזואליזציות האלו יחד הובילו למסקנה שהמשתנים ca, thalach, ו-oldpeak הם החזקים ביותר בחיזוי סיכון למחלת לב.

<u>הסיפור:</u> מחלות לב הן מהסיבות המובילות לתמותה עולמית, ולכן יש חשיבות מכרעת לחזות סיכון מוקדם.
באמצעות ניתוח נתונים שנאספו ממטופלים שונים, ניתן לזהות את המאפיינים הקליניים המרכזיים שמשפיעים על הסיכון למחלת לב. בניתוח זה, בדקנו את הביצועים של מספר מודלים חישוביים על נתוני המטופלים, ונמצא כי המודל Random Forest הציגו את הדיוק הגבוה ביותר בזיהוי סיכון למחלת לב. משתנים כמו גיל, לחץ דם וכאב חזה התגלו כמשפיעים מובהקים.

סיכום הנרטיב מאחורי סט הנתונים

סט הנתונים עוסק בניתוח נתוני מחלות לב. הוא מכיל מספר תכנים רלוונטיים לזיהוי ומניעת מחלות לב, ומספק מידע על תכונות שונות של חולים. הנה סיכום עיקרי התכנים והנרטיב:

<u>מטרות השימוש בנתונים:</u>

- ניבוי סיכון למחלות לב: הנתונים משמשים לניבוי הסיכון של אדם לפתח מחלת לב. לדוגמה, גיל, רמות כולסטרול, ולחץ דם הם גורמים חשובים בקביעת הסיכון.
- ניתוח התפלגות מאפיינים של החולים מאפשר להבין את השכיחות של תכונות מסוימות ולזהות דפוסים.
- השוואת ביצועים של מודלים: בעבודה עם נתונים אלו, ניתן להשוות בין מודלים שונים ולבחון את ביצועיהם בניבוי סיכון למחלה באמצעות כלי הערכה כמו ROC AUC.

<u>ניתוחים והדמיות:</u>

- תרשימים ותצוגות גרפיות: שימוש בגרפים כמו תרשימי עמודות, גרפי פיזור ו-boxplots כדי להמחיש את ההתפלגות והקשרים בין משתנים שונים.
- מודלים סטטיסטיים: יישום של מודלים סטטיסטיים ואלגוריתמים של למידת מכונה כדי לחזות את הסיכון למחלות לב ולבחון את ביצועי המודלים השונים.

<u>מסקנות:</u>

- השפעה של גורמים שונים: גיל, לחץ דם וכולסטרול הם גורמים חשובים בקביעת סיכון למחלת לב.
- התפלגות והבדלים: ניתוח ההתפלגות של משתנים כמו סוג כאב בחזה ומין יכול לחשוף הבדלים בין קבוצות אוכלוסיה ולסייע בהבנה טובה יותר של המאפיינים הקליניים של חולים עם מחלת לב.
- הערכת ביצועים של מודלים: השוואת ביצועים של מודלים שונים מאפשרת לבחור את המודל האופטימלי לניבוי סיכון מחלה.

שאלות שעולות להמשך חקירה:

- כיצד ניתן לייעל את הטיפול בחולים על סמך המידע?
- איך ניתן להשתמש בניתוח הנתונים כדי לשפר את אסטרטגיות הטיפול בחולים עם סיכון גבוה?
- האם יש גורמים נוספים שאינם כלולים בסט הנתונים הנוכחי, כמו עישון, היסטוריה משפחתית או פעילות גופנית, שיש להם השפעה משמעותית על הסיכון למחלות לב?